tarea 4 - muestreo

Valentina Gonzalez - Andres Felipe Beltran Rodriguez

12/19/2021

# Primer punto

Los datos anexos representan en una población de 200 cerdos los cuales fueron pesados en (libras) y posteriormente fueron sometidos a una dieta durante 30 días, con el objetivo de determinar si el tratamiento genera una ganancia de peso en los animales, los animales fueron medidos al final del estudio y se obtuvo X2, el peso final.

## a

Suponga que los 200 cerdos representan la población de interés. Seleccione una muestra aleatoria simple de tamaño 50 por el método de Fan Muller y estime con X1 el peso medio de la población, construya un intervalo de confianza del 90% para la media, calcule el coeficiente de variación estimado y concluya.

Primero cargamos los paquetes requeridos para la carga de datos y para hacer el muestreo:

library(readxl)  
library(TeachingSampling)  
library(sampling)

Luego cargamos los datos en la tabla de datos, TAREA4.xls en la hoja ejercicio1 saltando las tres primeras lineas:

Ej1 <- read\_excel("./TAREA4.xls",   
 sheet = "ejercicio1", skip = 3)

Luego fijamos la semilla para la generación de números pseudo aleatoreos con la fución set.seed():

La función S.SI realiza un muestreo aleatorio simple sin reemplazo con una muestra de tamaño con el metodo de selección-rechazo [(fan, Muller 1962)](https://www.tandfonline.com/doi/abs/10.1080/01621459.1962.10480667).

set.seed(1)  
seleccion<-S.SI(200,50)

Una vez tenemos el vector seleccion, podemos usarlo para extraer las posiciones asociadas a la muestra desde la población, Ej1$X1:

Muestra<-Ej1$X1[seleccion]

Al tener la muestra, podemos determinar el estimador de la media al sumar todos sus elementos y dividir por el número de elementos, length(Muestra) =

sum(Muestra)/length(Muestra)

## [1] 51.9

Al tomar una muestra aleatoria simple de 50 chanchos se puede inferir que el peso medio antes del tratamiento es de 51.9 kg

Ahora, podemos crear una función que nos permitirá estimar el intervalo de confianza para la media, la cual depende de los siguientes argumentos:

* N = número de elementos en la población
* n = número de elementos en la muestra
* media = la media estimada mediante los elementos de la muestra
* desv = la desviación estándar de la muestra
* PorcentajeConf = El porcentaje de confianza utilizado para el cálculo de los cuantiles, necesarios para calcular el intervalo de confianza.

Estimacion\_media <- function(N,n,media,desv,PorcentajeConf)   
 {   
   
 f <- n/N  
 alf <- (1-(PorcentajeConf/100))/2  
 alfmed <- qnorm(alf, mean = 0, sd = 1, lower.tail = F)  
 varmed <- (desv^2/n)\*(1-f)  
 desmed <- sqrt(varmed)  
 a1 <- media-alfmed\*desmed  
 b1 <- media+alfmed\*desmed  
 cat("media: IC = ",a1, "--",b1,"\n")  
 }

Podemos entonces utilizar la función Estimacion\_media con los argumentos asignados para este ejercicio:

* N = 200 cerditos totales
* n = 50 cerditos en la muestra
* media = mean(Muestra)
* desv = sqrt(var(Muestra))
* PorcentajeConf = 90

Estimacion\_media(200,50,mean(Muestra),sqrt(var(Muestra)),90)

## media: IC = 50.39178 -- 53.40822

Al tomar una muestra aleatoria simple de 50 marranos, es posible afirmar con un 90 de confianza que el peso medio de los marranos está entre 50.4 y 53.4 kg.

Para determinar el error de muestreo, o el coeficiente de variación estimado podemos también crear una función llamada CVE\_media, la cual depende de los siguientes argumentos:

* N = tamaño de la población
* n = tamaño de muestra
* muestras = un vector con los elementos de la muestra

CVE\_media <- function(N,n,muestras)   
 {   
 f=n/N  
 w=(sqrt(var(muestras)\*(1-f)/n))/mean(muestras)\*100  
 cat(w)  
 }

dado que ya tenemos la muestra seleccionada en el vector Muestra podemos pasarlo como argumento en la función CVE\_media:

CVE\_media(200,50,Muestra)

## 1.766729

El error de muestreo cuando se determina la masa media al seleccionar una muestra aleatoria simple de 50 marranos a partir de una población de 200 mediante el método de selección - rechazo descrito por [(fan, Muller 1962)](https://www.tandfonline.com/doi/abs/10.1080/01621459.1962.10480667) es del 1.77

## b

Con la misma muestra aleatoria simple de tamaño 50, estime con la variable X2 el peso medio de la población, construya un intervalo de confianza del 90% para la media, calcule el coeficiente de variación estimado y concluya. Podría afirmar que hubo ganancia en el peso?

Primero fijamos la semilla para obtener la misma muestra aleatoria simple:

set.seed(1)

Luego realizamos la extracción de los datos a partir de la variable X2:

seleccion1 <- S.SI(200,50)  
Muestra2<-Ej1$X2[seleccion1]

Una vez seleccionamos la muestra podemos estimar el valor de peso promedio:

sum(Muestra2)/50

## [1] 61.72

A su vez, podemos estimar un intervalo de confianza para la media, y el error de muestreo o coeficiente de variacion estimado:

Estimacion\_media(200,50,mean(Muestra2),sqrt(var(Muestra2)),90)

## media: IC = 59.47191 -- 63.96809

CVE\_media(200,50,Muestra2)

## 2.214423

Para la misma muestra aleatoria de marranos, se estima que la masa media es de 61.72 kg después del tratamiento.

Para saber si el tratamiento tuvo efecto, es decir, si los marranitos subieron de peso, podemos utilizar los resultados de masa para la muestra seleccionada antes y después del tratamiento, para esto podemos hacer una prueba de hipótesis con la función t.test:

t.test(Muestra,Muestra2)

##   
## Welch Two Sample t-test  
##   
## data: Muestra and Muestra2  
## t = -5.1672, df = 85.679, p-value = 1.53e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -13.598148 -6.041852  
## sample estimates:  
## mean of x mean of y   
## 51.90 61.72

Se puede inferir con un 95% de confianza que el tratamiento tuvo efecto en el engorde de los cerdos, ya que la diferencia entre las medias de la muestra seleccionada antes y después del tratamiento es diferente de cero.

1. Estime el parámetro del total bajo un 95% de confianza con la misma muestra X1, concluya. Construya un intervalo de confianza del 95% para la media, calcule el coeficiente de variación estimado y concluya.

sum(Muestra)/(50/200)

## [1] 10380

Estimacion\_total=function(N,n,desv,total,PorcentajeConf)   
{ f=n/N  
alf=(1-(PorcentajeConf/100))/2  
alfmed=qnorm(alf, mean = 0, sd = 1, lower.tail = F)  
 vartotal=N\*N\*(desv^2/n)\*(1-f)  
 destotal=sqrt(vartotal)  
 a2=total-alfmed\*destotal  
 b2=total+alfmed\*destotal  
 cat("total: IC = ",a2, "--",b2,"\n")  
}  
Estimacion\_total(200,50, sqrt(var(Muestra)),10380,95)

## total: IC = 10020.57 -- 10739.43

CVE\_total=function(N,n,muestras)   
{ f=n/N  
v= (sqrt(N\*N\*(var(muestras)/n)\*(1-f)))/(sum(muestras)/(50/200))\*100   
 cat(v)  
}  
  
CVE\_total(200,50,Muestra)

## 1.766729

Estimacion\_media(200,50,mean(Muestra),sqrt(var(Muestra)),95)

## media: IC = 50.10285 -- 53.69715

CVE\_media(200,50,Muestra)

## 1.766729

# Segundo punto

## a

Los datos anexos representan libras de grasa de leche producida para dos poblaciones independientes, donde se midieron las libras de grasa de leche para 100 vacas Holstein y 100 vacas Normando.

Seleccionar una muestra aleatoria de tamaño 60 de cada población por un muestreo aleatorio simple por el método de coordinado negativo. Estime la media de libras de grasa de leche para cada una de las poblaciones de interés, construya un intervalo de confianza del 95%, y calcule el coeficiente de variación estimado. Interpretar los resultados obtenidos.

Ej.2 <- read\_excel("./TAREA4.xls",   
 sheet = "ejercicio2", skip = 2)  
View(Ej.2)  
set.seed(8)  
seleccion1<-srswor(60,100)   
Muestra3<-Ej.2$X1[seleccion1==1]  
Muestra3.2<-Ej.2$X2[seleccion1==1]  
  
mean(Muestra3)

## [1] 41.46667

Estimacion\_media(100,60,mean(Muestra3),sqrt(var(Muestra3)),95 )

## media: IC = 39.49776 -- 43.43557

CVE\_media(100,60, Muestra3)

## 2.422577

mean(Muestra3.2)

## [1] 50.23333

Estimacion\_media(100,60,mean(Muestra3.2),sqrt(var(Muestra3.2)),95 )

## media: IC = 48.2043 -- 52.26237

CVE\_media(100,60,Muestra3.2)

## 2.060862

t.test(  
 x = Muestra3,  
 y = Muestra3.2,  
 alternative = "two.sided",  
 mu = 0,  
 conf.level = 0.95  
)

##   
## Welch Two Sample t-test  
##   
## data: Muestra3 and Muestra3.2  
## t = -3.8436, df = 117.89, p-value = 0.0001969  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -13.283365 -4.249969  
## sample estimates:  
## mean of x mean of y   
## 41.46667 50.23333

## b

Los datos anexos corresponden al precio de papa de dos variedades (Parda Pastusa y Diacol Capiro). Los datos medidos comprenden los años 1995 y 2011 en Colombia. Con los datos seleccionar una muestra aleatoria de tamaño 40 por el método de muestreo aleatorio sistemático y estimar el precio promedio de las dos variedades de papa y el porcentaje de error de muestreo cometido e interpretar los resultados obtenidos.

Ej2.b <- read\_excel("./TAREA4.xls",   
 sheet = "ejercicio2.1")  
round(60/40,0)

## [1] 2

set.seed(3)  
seleccion2<-S.SY(60,2)  
Seleccion2.1<-seleccion2  
ca<-seq(1, 20, 2)  
seleccion2.2<-c(Seleccion2.1,ca)  
Muestra4.1<-Ej2.b$rendpastusa  
Muestra4.1<-Muestra4.1[seleccion2.2]  
  
Muestra4.2<-Ej2.b$rendicapiro  
Muestra4.2<-Muestra4.2[seleccion2.2]  
  
mean(Muestra4.1)

## [1] 9625.158

Estimacion\_media(60,40,mean(Muestra4.1),sqrt(var(Muestra4.1)),95 )

## media: IC = 8783.67 -- 10466.65

CVE\_media(60,40, Muestra4.1)

## 4.460585

mean(Muestra4.2)

## [1] 18044.94

Estimacion\_media(60,40,mean(Muestra4.2),sqrt(var(Muestra4.2)),95 )

## media: IC = 17345.91 -- 18743.97

CVE\_media(60,40,Muestra4.2)

## 1.976468

t.test(  
 x = Muestra4.1,  
 y = Muestra4.2,  
 alternative = "two.sided",  
 mu = 0,  
 conf.level = 0.95  
)

##   
## Welch Two Sample t-test  
##   
## data: Muestra4.1 and Muestra4.2  
## t = -8.7094, df = 75.462, p-value = 5.076e-13  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -10345.45 -6494.12  
## sample estimates:  
## mean of x mean of y   
## 9625.158 18044.941

####Comparacion de medias ? -diferencia de medias corregir

# tercer punto

Los datos anexos representan una población de 120 plantas de arveja de dos variedades, las cuales fueron inoculadas para determinar si sobrevivían o no a la inoculación.

## a

Suponga que los datos representan la población. Seleccione una muestra aleatoria simple de tamaño 70 por medio de un muestreo sistemático, estime la proporción de plantas que sobreviven de la variedad 1. Construir el intervalo de confianza al 99% para hacer inferencia y el porcentaje de error de muestreo concluya.

Ej.3 <- read\_excel("./TAREA4.xls",   
 sheet = "ejercicio3", col\_types = c("numeric",   
 "numeric", "numeric", "numeric",   
 "numeric"), skip = 3)

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting numeric in E6 / R6C5: got 'PUNTO B'

## New names:  
## \* `` -> ...3  
## \* `` -> ...4  
## \* `` -> ...5

round(120/70,0)

## [1] 2

set.seed(4)  
seleccion3<-S.SY(120,2)  
caa<-seq(2, 20, 2)  
seleccion3.1<-c(seleccion3,caa)  
Muestra5<-Ej.3$`VARIEDAD 1`[seleccion3.1]  
sum(Muestra5)/70

## [1] 0.5142857

Estimacion\_proporcion=function(n,EstProporcion,PorcentajeConf)   
{   
 alf=(1-(PorcentajeConf/100))/2  
alfmed=qnorm(alf, mean = 0, sd = 1, lower.tail = F)  
 a3=EstProporcion-(alfmed\*sqrt((EstProporcion\*(1-EstProporcion ))/n) )  
 b3=EstProporcion+(alfmed\*sqrt((EstProporcion\*(1-EstProporcion ))/n))  
 cat("Proporción: IC = ",a3, "--",b3,"\n")  
}  
  
Estimacion\_proporcion(70,sum(Muestra5)/70,99)

## Proporción: IC = 0.3604133 -- 0.6681581

CVE\_proporcion=function(n,EstProporcion)   
{   
w= (sqrt((EstProporcion\*(1-EstProporcion ))/n))/EstProporcion  
 cat(w)  
}  
  
CVE\_proporcion(70,sum(Muestra5)/70)

## 0.1161553

Los datos anexos en la misma hoja representan una muestra aleatoria de 100 plantas de arroz las cuales fueron transformadas con un gen que protege potencialmente a las plantas de la salinidad. Posteriormente se le adicionaron 300 mm de Cloruro de Sodio, se observo si germino o no la planta. Se codifico como (1) si tolera la salinidad, (0) si no la tolera.

## b

Suponga que los datos representan la población. Seleccione una muestra aleatoria simple de tamaño 50 por medio de un muestreo sistemático, estime la proporción de plantas que no toleran la salinidad. Construya el intervalo de confianza, estime el error de muestreo e interprete.

Muestra5.5<-Ej.3[c(3:102),5]  
round(100/50,0)

## [1] 2

set.seed(5)  
seleccion4<-S.SY(100,2)  
seleccion4.1<-c(seleccion4)  
Muestra5.6<-Muestra5.5[seleccion4.1,]  
1-(sum(Muestra5.6)/50)

## [1] 0.48

Estimacion\_proporcion(50,1-(sum(Muestra5.6)/50),99)

## Proporción: IC = 0.2980071 -- 0.6619929

CVE\_proporcion(50,1-(sum(Muestra5.6)/50))

## 0.147196

# Cuarto punto

Se midieron tres especies de árboles (alnus acuminata Aa, acacia decurrens Ad, sambucus S). En cada una de ella se midieron en diferentes fechas durante 15 días la altura en centímetros y el diámetro basal en centímetros. Suponga que se va a realizar un muestreo estratificado y en cada estrato un muestreo aleatorio simple. Tome como variable de estratificación el tipo de especie y seleccione una muestra aleatoria simple de tamaño 70 y haga una asignación proporcional al tamaño del estrato, utilice el método coordinado negativo para la selección de la muestra en cada estrato, suponga que el total de elementos en cada estrato es la población. Suponga que los tamaños de los estratos son los que se forman.

Con la altura y el diámetro basal (cm) para el día 150 estime la media para las dos variables de interés en cada estrato y estime la media total para toda la población. Construya un intervalo de confianza del 95% para la media en cada estrato y para la media de toda la población, calcule el coeficiente de variación estimado en cada estrato y el de toda la población y concluya.

Ej.4 <- read\_excel("./TAREA4.xls",   
 sheet = "ejercicio4", na = ",")

## New names:  
## \* `` -> ...1  
## \* `` -> ...2  
## \* `` -> ...3  
## \* `` -> ...4  
## \* `` -> ...5  
## \* ...

View(Ej.4)  
Altura<-Ej.4[c(4:155),c(1,12)]  
colnames(Altura)<-c("Especie","Dia150")  
especies<-Altura$Especie  
Diametro<-Ej.4[c(161:312),c(1,12)]  
colnames(Diametro)<-c("Especie","Dia\_150")  
Diametro$Especie<-as.factor(Diametro$Especie)  
  
##Asignacion proporcional  
v1<-c("A.a","A.d","S")  
v2<-c(40,40,72)  
v3<-c(40\*100/152,40\*100/152,72\*100/152)  
v4<-c(40\*70/152,40\*70/152,72\*70/152)  
tabla1<-data.frame(v1,v2,v3,v4)  
colnames(tabla1)<-c("Especie","Nro de arboles", "Proporcion","Nro arboles en la muestra")

knitr::kable(tabla1,"simple")

|  |  |  |  |
| --- | --- | --- | --- |
| Especie | Nro de arboles | Proporcion | Nro arboles en la muestra |
| A.a | 40 | 26.31579 | 18.42105 |
| A.d | 40 | 26.31579 | 18.42105 |
| S | 72 | 47.36842 | 33.15789 |

###ALTURA  
  
estratos<-strata(Altura,stratanames = c("Especie"),size = c(18,18,34) ,method = "srswor",description = T)

## Stratum 1   
##   
## Population total and number of selected units: 40 18   
## Stratum 2   
##   
## Population total and number of selected units: 40 18   
## Stratum 3   
##   
## Population total and number of selected units: 72 34   
## Number of strata 3   
## Total number of selected units 70

altura.muestreado <- getdata( Altura, estratos )  
Alturaest1<-altura.muestreado[c(1:18),1]  
Alturaest2<-altura.muestreado[c(19:36),1]  
Alturaest3<-altura.muestreado[c(37:70),1]  
  
## Altura estrato 1  
mean(Alturaest1)

## [1] 94.83333

Estimacion\_media(40,18,mean(Alturaest1),sqrt(var(Alturaest1)),95)

## media: IC = 88.13075 -- 101.5359

CVE\_media(40,18,Alturaest1)

## 3.606063

## Altura estrato 2  
mean(Alturaest2)

## [1] 130.1111

Estimacion\_media(40,18,mean(Alturaest2),sqrt(var(Alturaest2)),95)

## media: IC = 111.8165 -- 148.4057

CVE\_media(40,18,Alturaest2)

## 7.173972

## Altura estrato 3  
mean(Alturaest3)

## [1] 103.7647

Estimacion\_media(72,34,mean(Alturaest3),sqrt(var(Alturaest3)),95)

## media: IC = 98.60903 -- 108.9204

CVE\_media(72,34,Alturaest3)

## 2.535056

### Altura total  
ta1<-sum(Alturaest1)/(18/40)  
ta2<-sum(Alturaest2)/(18/40)  
ta3<-sum(Alturaest3)/(34/72)  
Medtotal<-sum(ta1,ta2,ta3)/152  
Medtotal

## [1] 108.3476

vara1<-18\*18\*(1-(18/40))\*var(Alturaest1)/18  
vara2<-18\*18\*(1-(18/40))\*var(Alturaest2)/18  
vara3<-34\*34\*(1-(34/72))\*var(Alturaest3)/34  
varmed<-sum(vara1,vara2,vara3)/(152\*152)  
desmed=sqrt(varmed)  
alf=(1-(95/100))/2  
alfmed=qnorm(alf, mean = 0, sd = 1, lower.tail = F)  
a1=Medtotal-alfmed\*desmed  
b1=Medtotal+alfmed\*desmed  
cat("media: IC = ",a1, "--",b1,"\n")

## media: IC = 105.7682 -- 110.9271

CVE\_alt\_tot<-sqrt(varmed)/Medtotal\*100  
CVE\_alt\_tot

## [1] 1.214671

###ALTURA  
  
estratos1<-strata(Diametro,stratanames = c("Especie"),size = c(18,18,34) ,method = "srswor",description = T)

## Stratum 1   
##   
## Population total and number of selected units: 40 18   
## Stratum 2   
##   
## Population total and number of selected units: 40 18   
## Stratum 3   
##   
## Population total and number of selected units: 72 34   
## Number of strata 3   
## Total number of selected units 70

Diametro.muestreado <- getdata( Diametro, estratos1 )  
Diam\_est1<-altura.muestreado[c(1:18),1]  
Diam\_est2<-altura.muestreado[c(19:36),1]  
Diam\_est3<-altura.muestreado[c(37:70),1]  
  
## Altura estrato 1  
mean(Diam\_est1)

## [1] 94.83333

Estimacion\_media(40,18,mean(Diam\_est1),sqrt(var(Diam\_est1)),95)

## media: IC = 88.13075 -- 101.5359

CVE\_media(40,18,Diam\_est1)

## 3.606063

## Altura estrato 2  
mean(Diam\_est2)

## [1] 130.1111

Estimacion\_media(40,18,mean(Diam\_est2),sqrt(var(Diam\_est2)),95)

## media: IC = 111.8165 -- 148.4057

CVE\_media(40,18,Diam\_est2)

## 7.173972

## Altura estrato 3  
mean(Diam\_est3)

## [1] 103.7647

Estimacion\_media(72,34,mean(Diam\_est3),sqrt(var(Diam\_est3)),95)

## media: IC = 98.60903 -- 108.9204

CVE\_media(72,34,Diam\_est3)

## 2.535056

### Altura total  
td1<-sum(Diam\_est1)/(18/40)  
td2<-sum(Diam\_est2)/(18/40)  
td3<-sum(Diam\_est3)/(34/72)  
Medtotal1<-sum(td1,td2,td3)/152  
Medtotal1

## [1] 108.3476

vard1<-18\*18\*(1-(18/40))\*var(Diam\_est1)/18  
vard2<-18\*18\*(1-(18/40))\*var(Diam\_est2)/18  
vard3<-34\*34\*(1-(34/72))\*var(Diam\_est3)/34  
varmed1<-sum(vard1,vard2,vard3)/(152\*152)  
desmed1<-sqrt(varmed1)  
alf=(1-(95/100))/2  
alfmed=qnorm(alf, mean = 0, sd = 1, lower.tail = F)  
a2=Medtotal1-alfmed\*desmed1  
b2=Medtotal1+alfmed\*desmed1  
cat("media: IC = ",a2, "--",b2,"\n")

## media: IC = 105.7682 -- 110.9271

CVE\_diam\_tot<-sqrt(varmed1)/Medtotal1\*100  
CVE\_diam\_tot

## [1] 1.214671

# quinto punto

## a

Suponga que necesita hacer un estudio en un cultivo de flores y necesita determinar el tamaño de muestra apropiado para realizar un estudio en las plantas. Si en total se tienen 400 rosas de una variedad de interés y desea cometer un error máximo relativo del 10% bajo una confianza del 95%. Cuál es el número de rosas a estudiar. Tenga en cuenta que se tenía información auxiliar que identificaba en una muestra de 45 de estas rosas si tenían o no el diámetro apropiado y de estás 25 si alcanzaban el diámetro esperado. Concluya.

z<-qnorm((1-(95/100))/2, mean = 0, sd = 1, lower.tail = F)  
E<-4545450.1

## b

Suponga la misma información de a y adicionalmente un deff de 1,32. Calcule el tamaño de muestra apropiado.

En todos los casos construya los intervalos de confianza apropiados para hacer inferencia, el error de muestreo e interprete.